AD A063762

CRC 345

20000726135

(15)

NW

LEVEL

THE PREDICTION OF ATTRITION FROM MILITARY SERVICE

CENTER FOR NAVAL ANALYSES
1401 Wilson Boulevard
Arlington, Virginia 22209
Institute of Naval Studies

DDC FILE COPY

By: John T. Warner

September 1978

Prepared for:

OFFICE OF NAVAL RESEARCH
Department of the Navy
Arlington, Virginia 22217

OFFICE OF THE CHIEF OF NAVAL OPERATIONS (Op96)
Department of the Navy
Washington, D.C. 20350

02 034500.00

79 01 26 038

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER	(14) CRC-345
2. GOVT ACCESSION NO.	
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle)	(6) The Prediction of Attrition From Military Service
5. TYPE OF REPORT & PERIOD COVERED	
6. PERFORMING ORG. REPORT NUMBER	
7. AUTHOR(s)	(10) John T. Warner
8. CONTRACT OR GRANT NUMBER(s)	(15) N00014-76-C-0001
9. PERFORMING ORGANIZATION NAME AND ADDRESS	Center for Naval Analyses 1401 Wilson Boulevard Arlington, Virginia 22209
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS	
11. CONTROLLING OFFICE NAME AND ADDRESS	Office of Naval Research Department of the Navy Arlington, Virginia 22217
12. REPORT DATE	(11) Sep[redacted]78
13. NUMBER OF PAGES	23
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	Office of the Chief of Naval Operations (Op96) Department of the Navy Washington, D.C. 20350 (12) 30p.
15. SECURITY CLASS. (of this report)	Unclassified
15a. DECLASSIFICATION DOWNGRADING SCHEDULE	

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

This Research Contribution does not necessarily represent the opinion of the Department of the Navy.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

attrition, enlisted personnel, false negative rate, false positive rate, linear probability model, logit probability model, maximum likelihood, ordinary least squares, probability, regression analysis

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper provides a comparison of four statistical models for predicting first-year attrition from the Navy. The models compared are the individual linear probability model, the grouped linear probability model, the individual logit probability model and the grouped logit probability model. For different qualifying scores the models are compared in terms of their ability to discriminate between attriters and non-attriters. Their ability to predict the actual attrition rates within future entry cohorts is also compared.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

403 542

79 01 26 020

16

Center for Naval Analyses
an affiliate of the University of Rochester

1401 Wilson Boulevard   Arlington, Virginia 22209   703/524-9400

An Equal Opportunity Employer

17 November 1978

MEMORANDUM FOR DISTRIBUTION LIST

Subj: Center for Naval Analyses Research Contribution 345

Encl: (1) CRC 345, "The Prediction of Attrition from Military Service," by John T. Warner, September 1978

1. Enclosure (1) is forwarded as a matter of possible interest.

2. This Research Contribution compares the efficiency and accuracy of four statistical models for predicting discrete measures of military effectiveness, from the characteristics of applicants for military service. The question is important where large numbers of applicants, such as a yearly input, are involved and where computational costs are a consideration. The grouped logit model was found to be the best one. It was used in the Recruiting, Retention, and Reenlistment (R^3) to relate the characteristics of CY 1973 recruits to survival through three years of service.

3. Research Contributions are distributed for their potential value in other studies and analyses. They do not necessarily represent the opinion of the Department of the Navy.

Christopher Jehn

CHRISTOPHER JEHN
Director
Institute of Naval Studies

Distribution List
Reverse page

Subj: Center for Naval Analyses Research Contribution 345

DISTRIBUTION LIST

Department of the Navy

21A1	CINCLANTFLT	
21A2	CINCPACFLT	
21A3	CINCUSNAVEUR	
A1	ASST SECNAV (MRA&L)	
A2A	NAVCOMPT	
A2A	OPA	
A2A	ONR	
A5	CHNAVPERS	
A6	DC/S, Manpower, USMC	
B3	NAT'L DEF UNIV	
B3	ARMED FORCES STAFF COLLEGE	
FA34	HUMRESMANCEN, LANT	
FB44	HUMRESMANCEN, PAC	
FF30	NAVMMACLANT	
FF30	NAVMMACPAC	
FF38	USNA, Nimitz Library	
FF48	HUMRESMANCEN	
FH7	NAVMEDRSCHINSTITUTE	
FJ18	NAVPERSPROGSUPPACT	
FJ76	COMNAVCRUITCOM	
FKA6A16	NAVPERSRANDCEN	
FT1	CNET	
FT2	CNATRA	
FT5	CNTECHTRA	
FT73	NAVPGSCOL	
FT75	NAVWARCOL	
FT87	HUMRESMANSCOL	

OpNav:
Op-09R
Op-09BH
Op-96
Op-099
Op-01
Op-10
Op-29
Op-39
Op-59

Other
Ass't Secretary of Defense, Program Analysis & Evaluation (2 copies)
Ass't Secretary of Defense, Manpower, Reserve Affairs & Logistics
Defense Documentation Center
Department of the Army (Adj Gen'l)(6 copies)
Department of the Air Force (AFXOD)
Institute for Defense Analyses
Human Resource Research Organization
The Rand Corporation
System Development Corporation

TABLE OF CONTENTS

THE PREDICTION OF ATTRITION FROM MILITARY SERVICE

INTRODUCTION

Many situations arise where individuals must be classified into some category on the basis of observed characteristics. This classification problem is faced daily by college administrators, bank loan officers, and company employment managers. Applicants have to be classified as "successes" or "failures" on the basis of their observed characteristics. Thus, college applicants might be classified as successes or failures on the basis of factors such as SAT scores and high school ranking, loan applicants on the basis of income or net worth, and job applicants on the basis of past training and experience.

Beginning with the seminal work of Fisher (reference 1), the classification problem has been studied intensively in the statistics literature. The approaches to the classification problem may be separated into two general classes, those based on a linear probability model, and those based on some non-linear probability distribution such as the logistic or normal. In either approach, an equation for the probability of being a "success" is fit to observed data, and the fitted equation is used to predict the success chances of new applicants. Then, a critical success chance, or qualifying score is picked. New applicants whose predicted chances equal or exceed this qualifying score are classified as successes, while those whose predicted chances are lower are classified as failures. The optimal score for distinguishing between successes and failures depends upon the expected cost of misclassifying new applicants.

Despite extensive discussions of the relative efficiency of linear and non-linear models in the theoretical literature on classification (e.g., references 2 and 3), we have not found a detailed applied comparison of them. The purpose of this research contribution is to make such a comparison of these models when they are estimated with very large samples and used to classify other large cohorts of people.[1]

This work is an outgrowth of a study on attrition of first-term enlisted personnel from the U.S. Navy. With the advent of the all-volunteer force and higher pay scales for enlisted personnel, attrition (personnel leaving the Navy before completion of their first enlistment) is becoming more and more costly, and the Navy, as well as the other services, is under considerable pressure from Congress to reduce it. In the process of estimating equations for attrition probabilities that could be used for screening applicants with high chances of attrition, we had to answer the question of which empirical method gave the best discrimination between attriters and non-attriters.

---

[1] Nerlove and Press (reference 2) do provide an empirical application of these models, but they were concerned primarily with estimating probabilities rather than classification. Also, their work is based on fairly small samples relative to the ones utilized here.

Two linear probability models are compared with two non-linear probability models. The two linear models are the individual linear and grouped linear probability models, respectively. The two non-linear models, which are based on the logistic distribution, are the individual logit and the grouped logit models, respectively. The individual linear, grouped linear, and grouped logit models are all estimated by ordinary least squares (OLS) or generalized least squares (GLS) while the individual logit model is estimated by the method of maximum likelihood.

These four models are reviewed in detail. Theoretical reasons for expecting that the logit models will provide a better fit to the data are noted. Four models of first year attrition are estimated with a sample of 30,000 individuals from the cohort of 67,000 non-prior service males who enlisted in the U.S. Navy in CY 1973. Next, the ability of the fitted equations to discriminate between the attriters and the non-attriters in a separate sample of 30,000 individuals from the CY 1973 cohort is analyzed. In addition, we analyze the ability of grouped linear and grouped logit equations fit with all of the data from the CY 1973 cohort to discriminate between attriters and non-attriters in the CY 1974 cohort of non-prior service male enlistees. Finally, we examine the question of which CY 1973 equation gives the better prediction of attrition rates in the CY 1974 cohort.

METHODOLOGIES FOR PREDICTING THE PROBABILITY OF ATTRITION

This section discusses the existing methodologies for estimating attrition probabilities and discriminating between attriters and non-attriters. It begins with a review of the individual and grouped linear probability models. The equivalence between the individual linear probability model and the linear discriminant function is noted. Then, the two logit models are discussed. Both are consistent and asymptotically efficient and should, therefore, yield similar parameter estimates in large samples. This is an important point, since the estimation of the individual logit model is considerably more expensive in large samples. Theoretical reasons for believing that the logit models will provide a better fit to the data than the linear models are examined.

Linear Models

To begin with, let $X = (X_1 \ldots X_k)$ be a 1 x k vector of variables which determine the probability that an individual will be an attriter. Then $p(A|X)$ is the conditional probability that the individual will attrite given X. The problem is to estimate the relationship between $p(A|X)$ and X. One way to do this is to assume a simple linear relationship between $p(A|X)$ and X:

$$p(A|X) = X\beta \text{ where } \beta = \begin{bmatrix} \beta_k \\ \cdot \\ \cdot \\ \cdot \\ \beta_k \end{bmatrix} . \qquad (1)$$

Equation (1) is called the linear probability model. The parameters in the linear probability model can be estimated two ways.

The Individual Linear Probability Model

The individual linear probability model is estimated by assigning a value of 1 to attriters and 0 to non-attriters. This binary dependent variable is then regressed on X. Formally, the model to be estimated is given in (2):

$$Y = X\beta + \epsilon \text{ where } Y + \begin{bmatrix} 1 \\ \vdots \\ 1 \\ \hline 0 \\ \vdots \\ 0 \end{bmatrix} \quad . \tag{2}$$

Y is an n x 1 vector of observations which may be partitioned into an n_1 x 1 vector of ones representing the n_1 attriters in the sample, and an n_2 x 1 vector of zeros representing the n_2 non-attriters. X in (2) is an n x K matrix of observations on the independent variables. The well-known OLS estimator of (2) is shown in (3):

$$\hat{\beta} = (X'X)^{-1}X'Y \quad . \tag{3}$$

After computing $\hat{\beta}$, the probability that an individual with set of characteristics X_i will attrite is $\hat{p}_i = X_i\hat{\beta}$.

The linear model is appealing because of the computational ease of OLS and because of the ability of OLS to handle very large samples. On the other hand, it has been subject to criticism in the literature. A major criticism is that the individual linear model violates the constant variance assumption of OLS. The error term in (2) is binominal -- it can take on the value $-X\beta$ or the value $1 - X\beta$. For the ith observation, the variance of the error term ϵ_i is $X_i\beta(1-X_i\beta)$. Since the error term is heteroskedastic, the OLS estimator of β will not be the minimum variance linear estimator.[1]

Goldberger (reference 4) suggests the following solution to this problem. First, (2) is estimated by OLS and the weight $w_i = \sqrt{X_i\hat{\beta}(1-X_i\hat{\beta})}$ is computed for each individual

[1] See Goldberger (reference 4) for a discussion of the problem of heteroskedasticity.

in the sample. Then, each Y_i and X_i is weighted by $1/w_i$, and Y_i/w_i is regressed on X_i/w_i. This weighting procedure yields a model with a constant error variance, and the regression of Y_i/w_i on X_i/w_i gives the generalized least squares (GLS) estimator of β, which is the minimum variance unbiased estimator (reference 4, p. 250).

A second criticism of the linear probability model is that it does not restrict $\hat{p}$ to lie within the unit interval, although a $\hat{p}$ outside of this interval could not be interpreted as a probability estimate. In addition, Goldberger's procedure for correction for heteroskedasticity is invalidated when predictions outside of the unit interval are obtained. While the problem of prediction outside of the unit interval should diminish as the sample size increases,[1] we still encountered it in the empirical work reported below with a sample of 30,000 observations. Nerlove and Press (reference 2, pp. 54-55) discuss some work by Smith and Cicchetti (reference 5) on methods for handling inadmissible weights obtained in the Goldberger procedure. We adopted the one that uses .02 as the estimate of $\hat{p}$ for the cases where p was less than zero. While this procedure can be applied to get around the problem of negative weights in the GLS estimation of β, the problem of interpreting the resulting equation as a probability model still remains.

A third criticism of the individual linear probability model is not so serious as it first appears. It is often stated that since the error term in (2) is not normally distributed, tests of significance are not exact tests. Ladd (reference 6) shows that despite the binary form of the dependent variable in (2), the usual tests of significance are exact tests.

The (unweighted) individual linear model is proportional to the linear discriminant function (LDF) first proposed by Fisher (reference 1) in 1936 as a means of identifying binary group membership. The goal of LDF is to derive some linear combination of known characteristics, say $Z = \lambda' X$, from known data, and then use this linear combination to identify the group to which a new applicant belongs. For the <u>ith</u> new applicant, if $Z_i = \lambda' X_i$ is less than some critical value of Z, say Z_o, the individual is classified as a member of group 1 (say, attriter). Otherwise he is classified as a member of group 2 (say, non-attriter).

Beginning with the assumption that X values are distributed multivariate normal with mean vector μ and variance-covariance matrix Σ, the "best" LDF coefficients are those which maximize ϕ in equation (4):

$$\phi = \frac{[\lambda' (\mu_A - \mu_{NA})]^2}{\lambda' \Sigma \lambda} \quad . \tag{4}$$

[1] Nerlove and Press (reference 2) note that extreme sensitivity of OLS estimator of β to the sample in small samples.

In (4), μ_A is the vector of means of the X values for the attriters and μ_{NA} is the vector of means of X values for the non-attriters. Thus, the λ vector is chosen such that the ratio of the squared difference between the means of the two groups, $\lambda'\mu_A$ and $\lambda'\mu_{NA}$, to their variance, $\lambda'\Sigma\lambda$, is maximized. The λ vector that maximizes (4) is given in (5):

$$\lambda = \Sigma^{-1}(\mu_A - \mu_{NA}) \quad . \tag{5}$$

The mean vectors μ_A and μ_{NA} and the variance-covariance matrix Σ are unobservable. However, λ can be estimated by using the sample averages $\overline{X}_A$ and $\overline{X}_{NA}$ as estimates of μ_A and μ_{NA} and the sample variances and covariances of the X values to estimate Σ. Thus, λ is estimated by (6):

$$\hat{\lambda} = S^{-1}[\overline{X}_A - \overline{X}_{NA}] \quad . \tag{6}$$

Ladd (reference 6) has shown that the $\hat{\lambda}$ vector obtained from (6) is directly proportional to the regression coefficient vector $\hat{\beta}$ obtained from (3). This relationship is shown in (7):

$$\hat{\lambda} = \hat{\beta}\left(\frac{n-2}{ESS}\right) \quad . \tag{7}$$

ESS in (7) is the error sum of squares from the linear regression (3). Thus, using a linear discriminant function to assign individuals to group 1 (say, attriters) or group 2 (say, non-attriters) with a cutting score of Z_o is equivalent to assignment on the basis of the linear probability model with a cutting score of $p_o = \left(\frac{ESS}{n-2}\right) Z_o$.

The LDF procedure is not subject to quite the same criticisms as the individual linear probability model, even though the parameter estimates from the two procedures are proportional to one another. Since the fitted LDF is not used to predict probabilities, but only for classification, it is not subject to the criticism that it gives predicted probabilities outside of the unit interval. In addition, there is no problem of heteroskedasticity since the estimation procedure is not based on the assumption of OLS that the error term is normally distributed with constant variance.

Grouped Linear Probability Model

An alternative to the linear probability model based on individual observations is the grouped linear probability model. In this procedure, the observations are grouped into cells based on all possible combinations of the independent variables. Grouping is easy if all of the independent variables are categorical variables (e.g., race). If some of the variables are continuous (e.g., education level or age), they have to be broken up into a (reasonably small) number of intervals in order to group the data. The number of cells is the product, over the number of classifiers (e.g., race, age, education), of the number of intervals for each classifier. Thus, if there are five classifiers and 3 intervals for each classifier, there will be $3^5 = 243$ cells into which observations can fall.

Once the data are grouped, the proportion, $\hat{p}_j = a_j/n_j$, of n_j individuals in the jth cell who were attriters is computed. $\hat{p}_j$ is an estimate of the true probability p_j that individuals who fall into the jth cell will attrite. To estimate the grouped linear probability model, $\hat{p}_j$ is regressed on dummy variables representing the different levels of the classifiers.

One problem with a simple regression between $\hat{p}_j$ and X_j is that the error term in the regression $(p_j - \hat{p}_j)$ has a non-constant variance, and hence the OLS estimator of β is not a minimum variance estimator. The variance of the error term $(p_j - \hat{p}_j)$ is $p_j(1 - p_j)/n_j$ and is inversely related to the cell size n_j. This heteroskedasticity problem is handled by weighting each observation by the inverse of the estimated standard deviation error term, $\sqrt{\frac{n_j}{\hat{p}_j(1 - \hat{p}_j)}}$. The weighted regression between $\hat{p}_j \sqrt{\frac{n_j}{\hat{p}_j(1 - \hat{p}_j)}}$ and $X_j \sqrt{\frac{n_j}{\hat{p}_j(1 - \hat{p}_j)}}$ gives the minimum variance unbiased estimator of β.

Since the dependent variable in the grouped linear model is actually a rate rather than a 1 or 0 criterion variable, as in the individual linear model, the grouped linear model seems more in the spirit of a probability model. Since the dependent variable in the grouped linear model lies in the unit interval one would think that the predictions with the fitted model would also be more likely to lie in this interval. Unfortunately, we have found with a large sample that this is not necessarily the case.

Logistic Models

Because of their ease of application, the linear probability models are frequently employed in the literature, especially the individual linear model. However, there are reasons for suspecting that the linear probability model is a poor specification of $p(A|X)$. McFadden (reference 3, p. 374) notes that the (weighted) least squares estimator of β in (2) is very sensitive to specification error. Further, Cox (reference 7) and Day and Kerridge (reference 8) show that, under a variety of assumptions, $p(A|X)$ is logistic rather than linear.[1] The form of $p(A|X)$ for the logistic distribution is shown in (8).

$$p(A|X) = \frac{X\alpha}{1+e^{X\alpha}} \text{ where } \alpha = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_k \end{bmatrix} . \tag{8}$$

There are several ways to estimate the parameter vector in (8). If X is indeed multivariate normal, the best linear estimates of the α vector in (6) would be the LDF coefficients in (6). This follows since $X\hat{\alpha}$ is normally distributed if X is multivariate normal, and it was shown above that the $\hat{\lambda}$ vector in (7) is the best linear unbiased estimate of α when X is multivariate normal.[2] However, Halperin, Blackvelder, and Verter (reference 9) show that if X is not multivariate normal, the LDF estimator of α will not be consistent. Consistent, asymptotically efficient estimates of α may be obtained from either the grouped logit or individual logit procedures.

Grouped Logistic Model

With a large sample, α can be estimated using linear regression. The logistic probability function in (8) can be transformed into the following log-linear equation which may be estimated with OLS:

$$\ln\left(\frac{p}{1-p}\right) = X\alpha . \tag{9}$$

The dependent variable here is the logarithm of the odds of being an attriter. To estimate this equation, the data are grouped into cells just as in the grouped linear model. Then $\ln(\hat{p}_j/(1-\hat{p}_j))$ rather than $\hat{p}_j$ is used as the dependent variable in the regression.

---

[1] That $p(A|X)$ is logistic was originally derived from the assumption that X is multivariate normal, but the authors cited show that $p(A|X)$ is logistic for a variety of other conditions, including the case where all the independent variables are dichotomous.

[2] This implies that better estimates of attrition probabilities can be obtained by plugging the LDF coefficients in (6) into (8) than by converting LDF parameter estimates to individual linear probability model estimates via (7) and estimating attrition probabilities with a linear equation.

One problem is that the error term in this regression has the non-constant variance $\frac{1}{n_j p_j(1-p_j)}$. Weighted regression, where each observation is weighted by $\sqrt{n_j\hat{p}_j(1-\hat{p}_j)}$, yields the generalized least squares estimator of α . This grouped logit procedure, due to Berkson (reference 10), is known as the minimum logit chi-square method. Cox (reference 6) shows that under very general conditions this method yields consistent, asymptotically efficient estimates of α .

Individual Logistic Model

The logistic probability function in (8) is a non-linear equation which may also be estimated by the method of maximum likelihood. Maximum likelihood estimation of (8) was developed because the grouped logit procedure is inapplicable in small samples where many cells are empty or have only a few observations. As Nerlove and Press (reference 2, p. 60) state, the maximum likelihood procedure yields parameter estimates that have desirable small sample properties.

To estimate (8) by the method of maximum likelihood, the likelihood function is formed, and the α vector which maximizes the value of the likelihood function is found. Since individual observations are used, we call this model the individual logistic model. The likelihood function is:

$$L = \prod_{y_i=1} \frac{X\alpha}{1+e^{X\alpha}} \prod_{y_i=0} \frac{1}{1+e^{X\alpha}} \qquad (10)$$

Since (10) is not a simple linear expression, the α vector has to be estimated by a non-linear, iterative technique. Using the Newton-Raphson technique, the α vector is estimated as follows. The logarithm of the likelihood function L is computed, and then the partial derivative of lnL with respect to each α , $(\partial \ln L/\partial \alpha_i)$, is computed. Denote this k x 1 vector of partial derivatives by $\ell(\alpha)$. This vector is called the "score." The point at which $\ell(\alpha) = 0$ is called the "efficient score," since the likelihood function is maximized at this point.[1]

[1] The equations that make up the efficient score are similar to the normal equations in a linear regression, but are non-linear and cannot be solved analytically as can the normal equations.

Next, the k x k matrix of second partial derivatives of lnL with respect to α, $(\partial^2 \ln L/\partial\alpha_i \partial\alpha_j)$, is calculated. Denote this matrix by $\mathcal{L}'$. The vector α is then estimated iteratively as follows:[1]

$$\hat{\alpha}_m = \hat{\alpha}_{m-1} - [\mathcal{L}'(\alpha)_m]^{-1} \mathcal{L}(\alpha)_m \quad . \tag{11}$$

The m subscript refers to the m<u>th</u> iteration. $[\mathcal{L}'(\alpha)_m]^{-1}$ is the inverse of $\mathcal{L}'(\alpha)_m$. On each iteration, $[\mathcal{L}'(\alpha)_m]^{-1}$ and $\mathcal{L}(\alpha)_m$ are evaluated with the sample data. The best fit (i.e., the α vector such that $\mathcal{L}(\alpha) = 0$) is found when $[\mathcal{L}'(\alpha_m)]^{-1} \mathcal{L}(\alpha_m)$ converges to zero. The "start values" in the iteration process are the LDF coefficients in (6).

The ML estimate of α is normally distributed with asymptotic covariance matrix $[\mathcal{L}'(\alpha)_m]^{-1}$. Thus, a t-test of the significance of α_i is α_i/S_{ii} where S_{ii} is the square root of the i<u>th</u> diagonal element of $[\mathcal{L}'(\alpha)_m]^{-1}$.

---

[1] It was noted above that $\mathcal{L}(\alpha) = 0$ cannot be solved analytically for α. However, equation (11) for $\hat{\alpha}_m$ is derived as follows. If $\mathcal{L}(\alpha)$ is expanded in a Taylor series around the arbitrarily selected point α_o, then

$$\mathcal{L}(\alpha) = \mathcal{L}(\alpha_o) + (\alpha - \alpha_o)\mathcal{L}'(\alpha_o) + 1/2(\alpha - \alpha_o)^2 \mathcal{L}''(\alpha_o) \ldots$$

Ignoring $1/2(\alpha - \alpha_o)^2 \mathcal{L}''(\alpha_o)$ and other higher order terms, setting $\mathcal{L}(\alpha)$ equal to zero, and solving for α, we find that,

$$\alpha = \alpha_o - [\mathcal{L}'(\alpha_o)]^{-1} \mathcal{L}(\alpha_o) \quad .$$

This equation gives a value for α by expanding $\mathcal{L}(\alpha)$ around the arbitrary point α_o. The best fitting α, $\hat{\alpha}_m$, is found by iterating on α until $[\mathcal{L}'(\alpha_o)]^{-1} \mathcal{L}(\alpha_o)$ vanishes.

Finally, it is worthwhile to note the similarity between parameter estimates obtained from a model based on the logit distribution and those from a model based on the normal distribution. Instead of assuming a logit distribution, one could assume that attrition probabilities follow a normal distribution with unit variance ($\alpha^2 = 1$):

$$p = \int_{-\infty}^{X\gamma} \frac{e^{-1/2t^2}}{\sqrt{2\pi}}\, dt \qquad (12)$$

The parameters in (12) must be estimated by maximum likelihood. This model is called the probit model. While the logit model in (9) and the probit model in (12) look different, their cumulative distributions are very similar. The logit distribution has variance $\frac{\pi^2}{3}$. Therefore, if the α obtained from the logit estimation procedure is weighted by $\frac{\sqrt{3}}{\pi}$, it will be virtually the same as the γ obtained from the probit estimation. The logit and probit estimates differ only by the scale factor $\sqrt{\frac{3}{\pi}}$. Logit estimation is used more often than probit estimation, because the logit probability function is closed form (does not have an integral that must be evaluated) and is therefore much easier to estimate.

EMPIRICAL RESULTS

The four models discussed above were applied to data from the CY 1973 cohort of non-prior service male enlistees. The dependent variable was whether or not the individual was lost before the end of one year of service. The independent variables were years of education, mental ability as measured by the Armed Forces Qualification Test, marital status, age, and race. Education was split into three categories, less than 12 years, 12 years, and more than 12 years. Individuals were classified into five standard mental groups (I, II, IIIU, IIIL, and IV) on the basis of their AFQT scores. Age was split into three categories, less than 18 years, 18 or 19 years, and greater than 19 years. The various combinations of education level, mental ability, age, race, and marital status (3x5x3x2x2) give rise to 180 cells that individuals can fall into.

The CY 1973 cohort contained approximately 67,000 men. We divided the first 60,000 of them into 2 samples of 30,000 each (with 7,000 left over) by alternatively assigning individuals to an "A" sample and a "B" sample. Then, the four models described above were estimated with each sample of data. Splitting the cohort into samples of 30,000 was necessary for comparing the individual logit model with the other models,

because the maximum likelihood computer program used to estimate this model can accommodate a maximum of 30,000 observations. Even with 30,000 observations, 2.5 hours of computer time were required to estimate it.

After the four models were fit with each sample of data, the ability of each fitted equation to discriminate between the attriters and the non-attriters in the other (cross-validation) sample was examined. On the basis of qualifying scores ranging from 60 to 100, each individual in the cross-validation sample was classified as an attriter or non-attriter. Thus, if the qualifying score is 75, individuals who have lower survival chances are labeled attriters and individuals with equal or higher scores are labeled non-attriters. For scores ranging from 60 to 100, we examined: (1) the percentage of the cross-validation sample that would be selected, (2) the "hit" rate, or percent of sample correctly identified as either attriters or non-attriters, (3) the "false negative" rate, or percent of sample labeled as attriters who actually stayed, and (4) the "false positive" rate, or percent of sample labeled as non-attriters who actually left.

The Parameter Estimates

Table 1 shows the parameter estimates obtained by applying the four procedures described above to one of the samples. Estimates obtained with the second sample are contained in appendix A. The estimates shown in the column labeled "Individual Linear" are those obtained with the weighted regression procedure described in the last section.[1] Table 1 also shows the LDF coefficients, which are proportional to the unweighted estimates (not shown) of the individual linear probability model.

Several conclusions are apparent from table 1. With the large sample used here, each of the two grouped models gives virtually the same fitted equation as its individual counterpart. Especially in the case of the two linear models, the parameter estimates obtained with the grouped linear model are in most cases the same down to the third decimal place as those obtained with the individual linear model. Differences in the predicted attrition probabilities obtained with the two linear equations are quite small. Although not as obvious, the differences in the estimates from the two logit models also imply trivial differences in estimated attrition probabilities.[2] The parameter estimates from either

[1] It was noted above that the unweighted estimates of the individual linear probability model gave predicted attrition chances of less than zero in some cases. This occurred for individuals who had more than 12 years of education and who were in mental group I. These individuals made up about 2 percent of the sample. The problem of negative weights in the weighted regression procedure was handled by assigning these individuals an attrition probability of .02 .

[2] A difference in a parameter estimate between the two procedures of about .10 will imply a difference in the predicted attrition probability of about .01 . Most of the differences between the parameter estimates obtained with the two logit procedures are considerably smaller than .10 .

TABLE 1

ESTIMATES OF PARAMETER VALUES,
SAMPLE FROM CY 1973 COHORT

Variable	Individual linear	Grouped linear	Individual logit	Grouped logit	Linear discriminant function
Ed < 12	-.105(17.04)	-.109(14.14)	-.672(21.23)[a]	-.656(14.42)	-.783(19.05)
Ed > 12	.028(3.88)	.032(3.79	.349(4.51)	.284(2.87)	.249(3.82)
Mental group I	.034(9.95)	.084(9.65)	1.179(9.32)	1.040(6.00)	.615(7.33)
Mental group II	.021(3.96)	.020(3.09)	.201(4.50)	.208(3.60)	.154(3.61)
Mental group IIIL	-.053(7.70)	-.052(6.20)	-.345(7.71)	-.342(6.00)	-.390(8.01)
Mental group IV	-.098(12.46)	-.097(10.04)	-.581(12.98)	-.571(9.75)	-.721(13.69)
Dependents	-.046(4.82)	-.039(3.61)	-.349(5.52)	-.403(5.21)	-.371(5.57)
Age < 18	-.031(4.16)	-.024(2.56)	-.145(3.24)	-.166(3.14)	-.168(3.38)
Age > 19	-.027(4.30)	-.022(3.51)	-.185(4.13)	-.169(3.24)	-.173(4.16)
Race (non-Caucasian)	.027(3.61)	.037(4.15)	.136(3.04)	.081(1.28)	.179(3.27)
Constant	.881(25.70)	.882(20.79)	1.959(61.96)	1.950(40.87)	2.076(23.71)
N	30,000	137	30,000	137	30,000

[a]"t" values in parentheses.

of the two logit procedures differ from the LDF coefficients on precisely the variables that have the most impact on the probability of attrition, the education and mental group variables. For most of the other variables, the deviations of the logit coefficients from the LDF coefficients are small.

The close correspondence between the parameter estimates from the two logit models is to be expected, since both have been shown in the theoretical literature to be consistent and asymptotically efficient. The similarity of results is important, because with large samples the individual logit model is considerably more expensive to estimate.

The Selection Ratio and Distributions of Correct and Incorrect Predictions for Linear and Logit Models

Figures 1 through 4 show the selection ratio, hit rate, false positive rate, and false negative rate for the individual linear and logit models for qualifying scores ranging from 60 to 100. While the distributions in these figures are based on the individual linear and logit models, the distributions obtained with the grouped models were virtually the same. The differences that exist between models are between the linear models and the logit models, not between the two versions of the same model.

Looking first at figure 1, at a qualifying score of 60, all of the cross-validation sample would be admitted. As the qualifying score is raised, obviously fewer people are selected. The sizeable differences between the selection ratios implied by the two methods occurs in the range of qualifying scores between 74 and 82. In this range, a higher percentage of the cross-validation sample would be selected with the logit model than with the linear model. The maximum difference between models occurs at a qualifying score of 79, where 5 percent more people would be selected using the logit model.

The hit rate distribution is shown in figure 2. Again, the range where sizeable differences in hit rates occur lies between cutting scores of 74 and 82. In this range, the logit model gives a somewhat higher hit rate than does the linear model. Again, the maximum difference occurs at a qualifying score of 79. Here the logit model gives a 3 percent higher hit rate than the linear model.

Figures 3 and 4 show the rate of false positive predictions and the rate of false negative predictions for the linear and logit models. As figure 5 shows the rate of false positive predictions declines as the qualifying score is raised, because only those whose survival chances exceed the higher qualifying score are selected. Again, most of the differences between the models occur in the range 74 to 82. At the qualifying score of 79, the logit model gives a one percent higher rate of false positives than the linear model. The higher rate of false positives for the logit model in the range 74 to 82 is due to the fact that in this range, a higher percentage of the applicant cohort would be enlisted using the logit model (recall figure 1).

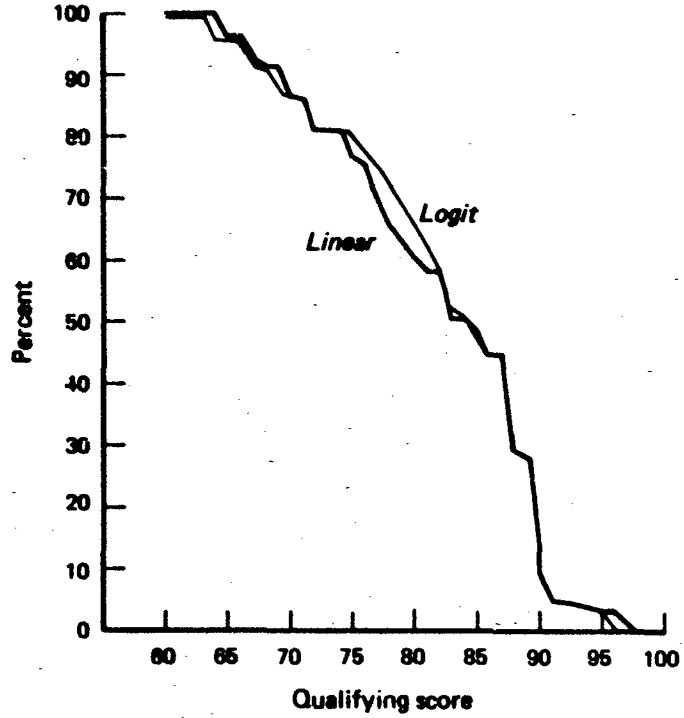

FIG. 1: SELECTION RATIO — CY73 A SAMPLE
EQUATIONS ON B SAMPLE

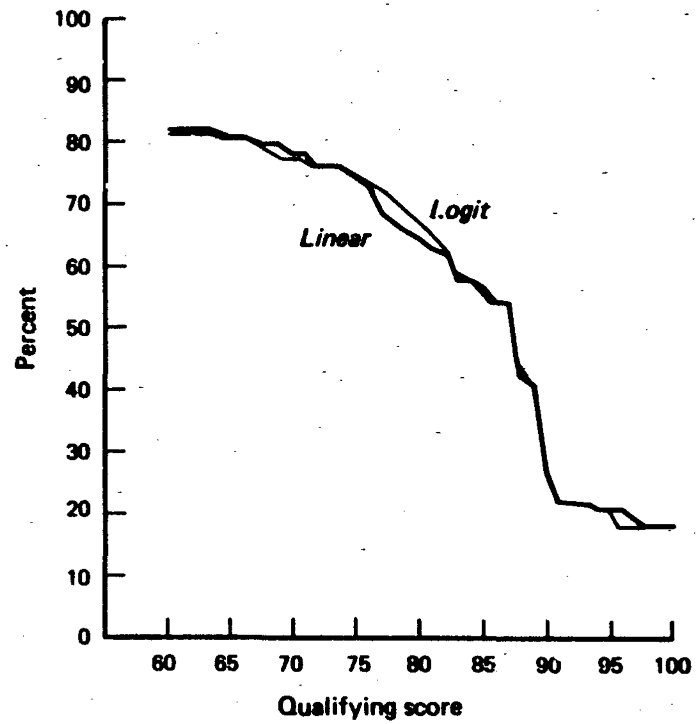

FIG. 2: HIT RATE — CY73 A SAMPLE
EQUATIONS ON B SAMPLE

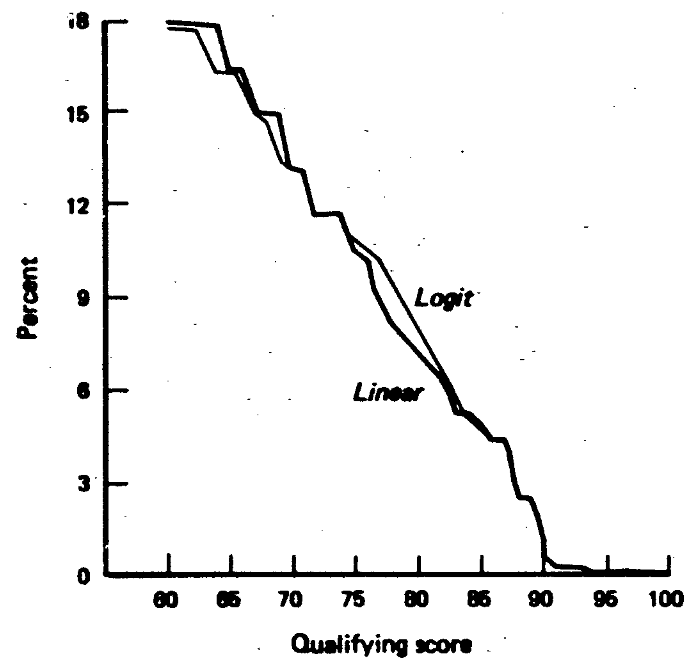

FIG. 3: FALSE POSITIVE RATE — CY73 A SAMPLE EQUATIONS ON B SAMPLE

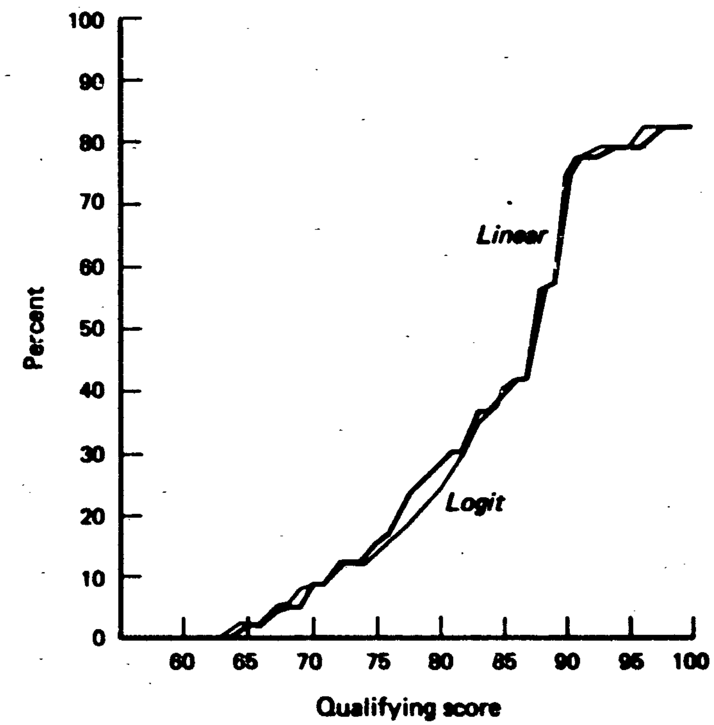

FIG. 4: FALSE NEGATIVE RATE — CY73 A SAMPLE EQUATIONS ON B SAMPLE

Looking at the rate of false negative predictions in figure 4, we see again, that, the differences between the logit and linear models occur in the range 74 to 82. Again, the maximum difference occurs at the qualifying score of 79, where the logit model has a 4 percent lower false negative rate than the linear model.

To summarize, our results indicate that individuals who have very low or very high chances of early attrition will be correctly classified by either model. Thus, for qualifying scores below 74 or above 82, the two models give about the same rate of hits, false positives, and false negatives. However, in the range between 74 and 82, the logit model gives a higher rate of hits and false negatives, but a higher rate of false positives. It is significant that this is the area of greatest overlap between attriters and non-attriters. Seventy-eight is the average SCREEN score of attriters, while 82 is the average score of non-attriters. In the range where overlap occurs, the logit model gives slightly better discrimination between attriters and non-attriters than the linear model.

Grouped logit and grouped linear equations fit to the whole CY 1973 cohort are found in Lockman (reference 11). These equations are reproduced in appendix B. The Navy is now using tables based on the grouped logit model to screen recruits, so we wanted to determine how well these two models distinguish between attriters and non-attriters in another cohort. Therefore, these equations were applied to the CY 1974 cohort. The selection ratio, the hit rate, false positive rate, and false negative rate distributions are shown in figures 5, 6, 7, and 8, respectively. Although the patterns are similar to the ones shown previously, the differences between the logit and linear models are much less pronounced. Whereas we found virtually no differences in the lower tails of the distributions in figures 1 through 4 above, we do find some differences in figures 5 through 8.

Prediction of Attrition Rates with Linear and Logit Models

In addition to using the linear or logit models for classification, we are also interested in just how well they predict future attrition rates. Even if the models are not used for recruit screening purposes, they could still be used to predict the attrition that will be suffered. As noted above, theory tells us that the logit model is a better specification of $P(A|X)$ than the linear model. If so, the logit model should have smaller errors in predicting future attrition rates than the linear model.

To see if this is true, we predicted the attrition rates for the 137 cells in the CY 1974 cohort which contained observations from grouped linear and grouped logit equations based on all of the data from the CY 1973 cohort. We computed two values reflecting the predictive ability of the two equations. The first is an error sum of squares, $\Sigma (p_j - \hat{p}_j)^2$. The second is an error sum of squares which weights the square of the error by the number of observations in the cell, $\Sigma N_j(p_j - \hat{p}_j)^2$. This statistic should provide a better comparison of prediction errors for two reasons. First, it weights each error by the "cost" of the error;

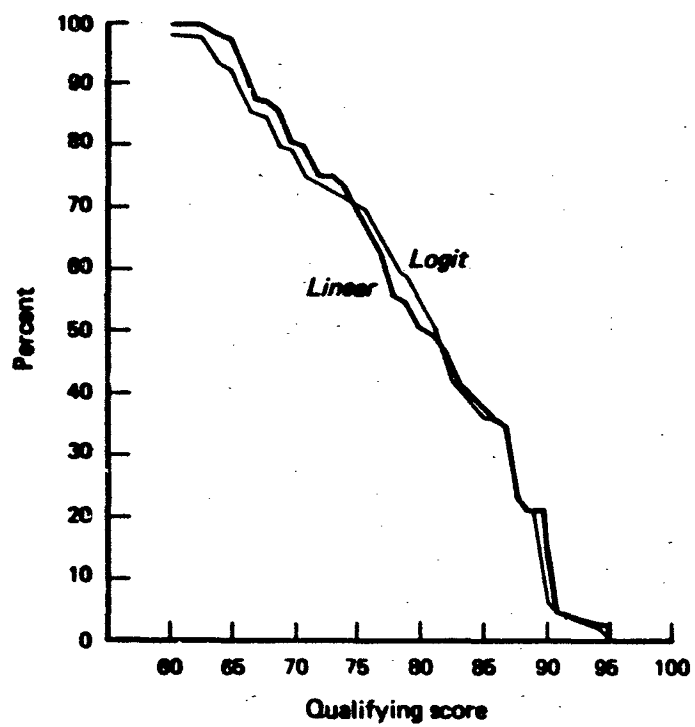

FIG. 5: SELECTION RATIO – CY73 EQUATIONS ON CY74 COHORT

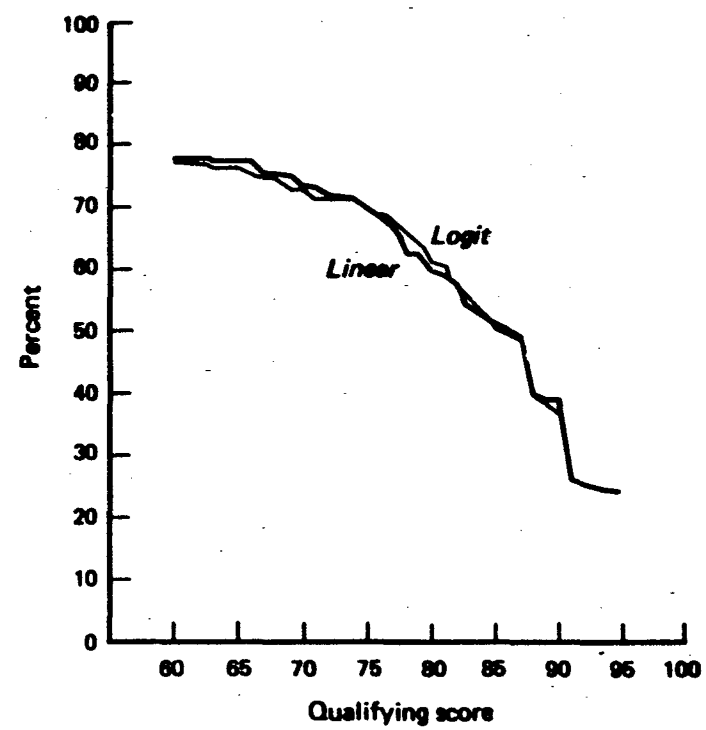

FIG. 6: HIT RATE – CY73 EQUATIONS ON CY74 COHORT

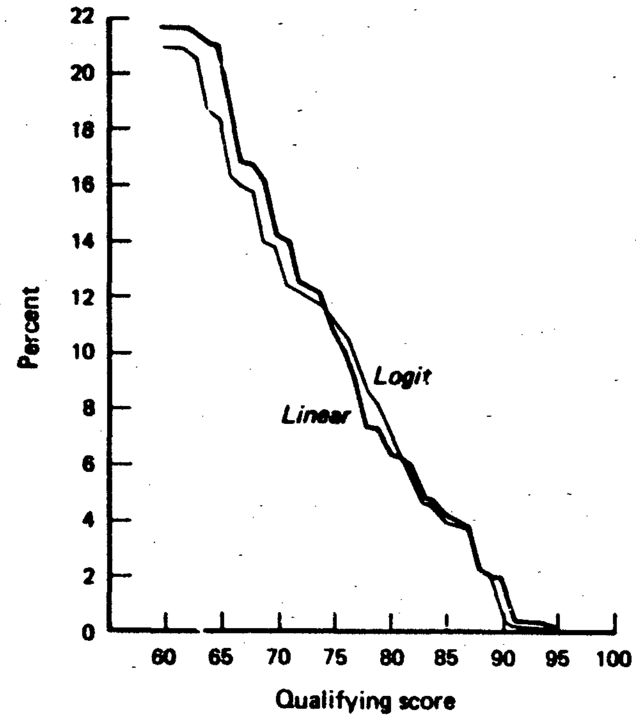

FIG. 7: FALSE POSITIVE RATE – CY73 EQUATIONS ON CY74 COHORT

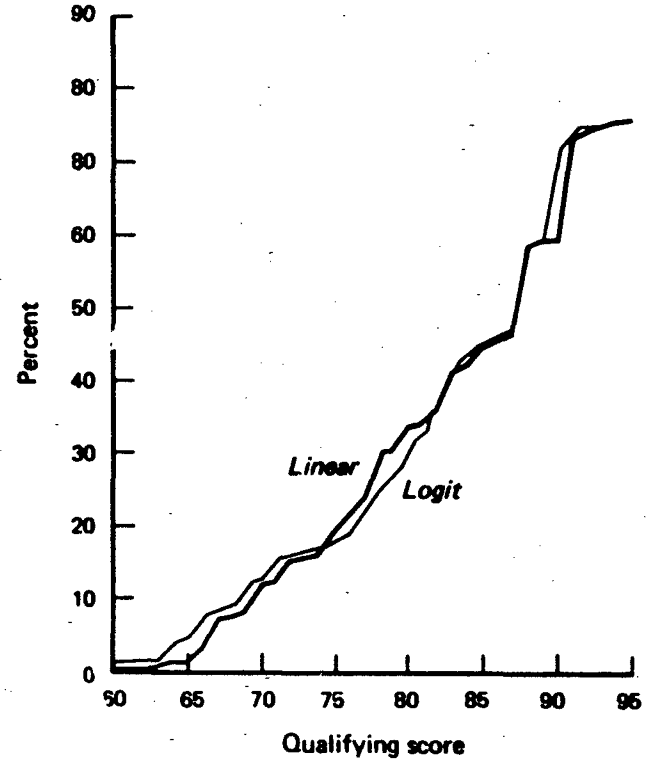

FIG. 8: FALSE NEGATIVE RATE – CY73 EQUATIONS ON CY74 COHORT

prediction errors are more expensive the more individuals there are in a cell. Second, it could be that the larger errors are occurring in the upper or lower tails of the probability distribution where the cell sizes are small. The linear model might be predicting as well in the middle of the distribution (say, where the attrition probabilities lie between .1 and .3) but doing poorly in the tails. If this is so, the difference between the models in $\Sigma N_j(p_j - \hat{p}_j)^2$ should be smaller than the differences in $\Sigma (p_j - \hat{p}_j)^2$.

Table 2 presents the results of these computations. Based on either measure, the logit model is found to give smaller prediction errors than the linear model. It is somewhat surprising that there is a larger (percentage) difference in the weighted error sum of squares between the models than in the unweighted error sum of squares. Differences between methods were not just due to the linear model having larger prediction errors in small cells or cells at the extremes of the probability distribution. These results indicate that the logit specification of $P(A|X)$ was a better predictor than the linear specification of $P(A|X)$.

TABLE 2

TWO MEASURES OF PREDICTION ERROR, CY 1973 EQUATIONS APPLIED TO CY 1974 COHORT

	Grouped logit	Grouped linear
$\Sigma (p_j - \hat{p}_j)^2$	4.08	4.38
$\Sigma N_j(p_j - \hat{p}_j)^2$	165.15	205.76

CONCLUSIONS

Several general conclusions emerge from our empirical analysis. First, with large samples, the individual linear and logit models give virtually the same fitted equation as their grouped counterpart. This is essentially an empirical demonstration of the fact that each individual model has the same asymptotic properties as its grouped counterpart. Knowing that the grouped logit model based on linear regression and the individual logit model based on maximum likelihood yield the same fitted equation is extremely useful, because maximum likelihood estimation is computationally expensive in very large samples.

Second, the logit models are found to be superior to the linear models on several counts. For a range of qualifying scores most likely to be used by the Navy to separate acceptable from unacceptable applicants, the logit models give somewhat better prediction of actual success or failure. They are also found to give a lower rate of "false negatives"

(predicted failures who are actual successes). However, they do give a slightly higher rate of "false positives" (predicted successes who are actual failures).

Third, the grouped logit model based on data from one cohort (CY 1973 enlistees) was found to give better estimates of attrition rates of different groups in other cohort (CY 1974 enlistees) than the grouped linear model based on the same data. Goodness of fit was measured by both weighted and unweighted error sums of squares in prediction. Consequently, the grouped logit model is the best model for the prediction of attrition with very large samples.

REFERENCES

1. Fisher, R.A., The Use of Multiple Measurement in Taxonomic Problems. Annals of Eugenics, Vol. 7, 1936, pp. 179-188.

2. Nerlove, M. and J.S. Press, Univariate and Multivariate Log-Linear and Logistic Models, Report R-1306-EDA/NIH, The RAND Corporation, Dec. 1973.

3. McFadden, D., Quantal Choice Analysis: A Survey, Annals of Economic and Social Measurement, Vol. 7, Fall 1976, pp. 363-390.

4. Goldberger, A., Economic Theory, John Wiley and Sons, New York, 1964.

5. Smith, V.K., and C.J. Cicchetti, "Estimation of Linear Probability Models with Dichotomous Dependent Variables," Resources for the Future 1972 (memo).

6. Ladd, B., "Linear Probability Functions and Discriminant Functions," Econometrica, Vol. 34, 1966, pp. 873-885.

7. Cox, D., Analysis of Binary Data, Methuen, London, 1970.

8. Day, N.E., and D.F. Kerridge, "A General Maximum Likelihood Discriminant," Biometrics, Vol. 23, 1957, pp. 313-323.

9. Halperin, M., Blackwelder, W.D., and J.I. Verter, "Estimation of the Multivariate Logistic Risk Function: A Comparison of the Discriminant and Maximum Likelihood Approaches," Journal of Chronic Diseases, Vol. 24, 1971, pp. 125-158.

10. Berkson, J., "Maximum Likelihood and Minimum Chi-Square Estimates of the Logistic Function," Journal of the American Statistical Association, Vol. 50, March 1955, pp. 130-162.

11. Center for Naval Analyses, Study 1068, "Chances of Surviving the First Year of Service: A New Technique for Use in Making Recruiting Policy and Screening Applicants for the Navy," by R.F. Lockman, Unclassified, November 1975.

APPENDIX A

ESTIMATES OF PARAMETER VALUES, SAMPLE B FROM CY 1973 COHORT

APPENDIX A

ESTIMATES OF PARAMETER VALUES, SAMPLE B FROM CY 1973 COHORT

Variable	Individual linear	Grouped linear	Individual logit	Grouped logit	Linear discriminant function
Ed < 12	-.112 (18.01)	-.117 (15.78)	-.727 (22.99)[a]	-.713 (16.45)	-.849 (20.92)
Ed > 12	.017 (2.32)	.025 (2.88)	.252 (3.25)	.201 (2.13)	.181 (2.74)
Mental Group I	.060 (7.79)	.064 (7.20)	.848 (6.70)	.752 (5.04)	.460 (5.50)
Mental Group II	.017 (3.34)	.018 (2.90)	.199 (4.47)	.197 (3.62)	.140 (3.32)
Mental Group IIIL	-.075 (10.64)	-.075 (9.08)	-.484 (10.83)	-.480 (8.85)	-.551 (4.37)
Mental Group IV	-.110 (13.95)	-.111 (12.11)	-.652 (14.58)	-.642 (11.56)	-.794 (15.38)
Marital status (married)	-.061 (6.21)	-.042 (3.94)	-.435 (6.87)	-.487 (6.67)	-.470 (7.10)
Age < 18	-.026 (3.43)	-.014 (1.55)	-.078 (1.75)	-.104 (2.05)	-.083 (1.67)
Age > 19	-.017 (3.18)	-.020 (3.28)	-.148 (3.31)	-.135 (2.70)	-.134 (3.25)
Race (Non-Caucasian)	.013 (1.67)	.022 (2.48)	.075 (1.68)	.035 (.67)	.113 (2.10)
Constant	.890 (24.20)	.890 (20.74)	2.017	2.101 (43.21)	2.141 (22.16)
N	30,000	131	30,000	131	30,000

[a]"t" values in parentheses.

APPENDIX B

PARAMETER ESTIMATES FOR GROUPED LOGIT AND GROUPED LINEAR MODELS, BASED ON TOTAL CY 1973 COHORT

APPENDIX B

PARAMETER ESTIMATES FOR GROUPED LOGIT AND GROUPED LINEAR MODELS, BASED ON TOTAL CY 1973 COHORT

Variable	Grouped logit	Grouped linear
Ed < 12	-.701 (21.20)[a]	-.111 (19.03)
Ed > 12	.314 (4.42)	.031 (4.49)
Mental Group I	.989 (8.37)	.079 (10.85)
Mental Group II	.254 (6.22)	.026 (5.28)
Mental Group IIIL	-.365 (8.85)	-.052 (7.91)
Mental Group IV	-.597 (14.23)	-.100 (13.44)
Marital status (married)	-.389 (6.95)	-.038 (4.36)
Age < 18	-.093 (2.76)	-.015 (2.89)
Age > 19	-.280 (6.43)	-.032 (5.43)
Race (Non-Caucasian)	.119 (2.64)	.034 (4.89)
Constant	1.976 (57.35)	.882 (26.89)
N	137	137

[a]"t" values in parentheses.